DOCUMENT RESUME

ED 306 280 TM 013 131

AUTHOR Ackerman, Terry A.; Davey, Tim C.
TITLE An Analysis of CAAP Essay and Multiple-Choice Writing Tests.
PUB DATE Mar 89
NOTE 26p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Entrance Examinations; *Essay Tests; Higher Education; Latent Trait Theory; *Multiple Choice Tests; Test Reliability; *Writing Evaluation
IDENTIFIERS *Collegiate Assessment of Academic Proficiency; Direct Assessment; Information Function (Tests); *Writing Tests

ABSTRACT

This study examines differences and similarities in the information provided by direct and indirect measures of writing from the Collegiate Assessment of Academic Proficiency (CAAP). The indirect measure was a 72-item multiple-choice test, while the direct measure involved responding to two essay prompts. The 40-minute multiple-choice test can be subdivided into six skill areas: (1) punctuation; (2) grammar and usage; (3) sentence structure; (4) style; (5) strategy; and (6) organization. Item response and essay ratings were calibrated together using a graded response model from item response theory. Results suggest that while the essays are measuring a different component of writing ability than the multiple-choice test, their overlap is substantial. Relative information plots also suggest that the writing sample provides information equal to as many as 40 multiple-choice items. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED306280

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Terry A. Ackerman

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

An Analysis of the CAAP Essay and Multiple-choice Writing Tests

Terry A. Ackerman

Tim C. Davey

The American College Testing Program

Paper presented at the 1989 AERA Annual Meeting, March 28, San Francisco.

Running Head: CAAP Writing Tests

TM013131

Abstract

This study examines differences and similarities in the information provided by direct and indirect measures of writing from the Collegiate Assessment of Academic Proficiency (CAAP). The indirect measure was a 72 item multiple-choice test, while the direct measure involved responding to two essay prompts. Item responses and essay ratings were calibrated together using a graded response model from item response theory. Results suggest that while the essays are measuring a different component of writing ability than the MC test, their overlap is substantial. Relative information plots also suggest that the writing sample provides information equal to as many as 40 multiple choice items.

A Comparison of the CAAP Essay and Multiple-choice Writing Tests

The American College Testing Program recently began development of the Collegiate Assessment of Academic Proficiency (CAAP), a battery of tests designed to measure selected academic skills which are foundational for performance in upper-level college courses and are essential to the general academic curriculum. CAAP tests measure skills in four areas: reading, mathematics, critical thinking and writing. Writing is measured both by multiple choice (MC) and by essay tests. The differences and the similarities in the information provided by these two test formats are the focuses of this paper.

Background

Experts in the area of writing assessment have long been divided as to the most appropriate way to assess writing skill. One belief is that the most efficient and most reliable method is to test using an MC format (Breland 1977). A second opinion is that essay tests are the best method because they are more "ecologically valid" (Braddock, Lloyd-Jones, & Schoer, 1963; Cooper & Odell, 1977). A third opinion is that each measures a different component of writing skill, and that whenever possible, both methods should be used (Ward, Frederiksen, & Carlson 1980; Ackerman & Smith, 1988).

Past research has focused on identifying the unique components that each type of assessment measures. For example, Ackerman and Smith (1988) theorized that writing skills can be characterized by a hierarchical continuum. MC tests, they maintained, tend to measure the ability to recognize proper language usage, whereas essay tests tend to measure not only recognition skills but also the ability to generate and organize written discourse. Ackerman and Smith factor analyzed results from three different measures of this continuum, an MC test, a free response test, and an essay test. They found three distinct factors, one for each test type.

The present study takes a completely opposite approach. That is, the purpose of this study is not to reestablish the uniqueness of essay and MC tests, but rather to determine the amount of common information provided.

It is acknowledged that essay and MC items indeed measure different components of writing, and that the different components measured are, in part, due to format effects. In fact, item response data, even from the same format, are almost always multidimensional to some extent. However, it is assumed that most multidimensional tests tend to measure a clearly dominant latent trait and several less explicit traits. Practitioners tend to view essay results and MC writing test results in this vein, believing that although the results from both measures may be different, there is a general dominant trait of writing skill that is being measured by both.

This can be illustrated briefly with a two-dimensional example. Consider the vector A shown in Figure 1. The length and direction of A represent the amount of discrimination along a composite of two abilities being measured by a hypothetical item. The long diagonal solid line represents a "reference composite"; or that composite of abilities being measured by the entire test. The length of vector B represents the estimated discrimination of item A mapped onto the reference composite. The closer the direction of the item vector to the direction of the reference composite, the smaller the loss in discrimination.

Insert Figure 1 about here

The CAAP Test

For field test purposes the CAAP essay test is composed of two essay prompts. Each prompt requires the student to read a passage and then, given a

specific context, to write an essay that argues a particular point of view about an issue drawn from the reading passage. Students are given 20 minutes to write on each prompt. Prompts were specifically developed to facilitate a "purpose-based" scoring system. An example essay prompt and a description of the scoring guide are provided in the Appendix. In this initial phase of development, each essay was scored by two independent raters (from a pool of about 40 raters) for general purpose (GP) and for language usage (LU). The GP rating reflects how well the examinees responded to the task required by the situations described in the prompts. The LU rating reflected the raters' impressions of the relative presence of usage or mechanical errors and the degree to which such errors impeded the flow of thought in the essays. The two rating of each characteristic are then averaged to provide single scores. If the ratings differ by more than a single point, a third rater arbitrates. Thus each student received one GP score and one LU score for each essay.

The 40-minute multiple-choice test has a 72-item test that can be subdivided into six skill areas: punctuation (abbreviated P containing 6 items), grammar and usage (G, 8), sentence structure (SS, 18), style (SL, 14), strategy (SA, 16), and organization (OR, 10).

Method

To assess whether or not the essay and MC tests are measuring similar or different skills, a factor analysis of the Pearson product-moment correlation matrix of the six MC content areas and the four essay ratings (a purpose rating and language usage rating for each of two essay prompts) was performed. This correlation matrix was corrected for unreliability.

To compare the amount of common information across tests parameter estimates were obtained using a graded response IRT model (Samejima, 1969).

In this model the probability for ordered responses r = k, k = 1, ... m; where response m reflects the highest rating value is given by:

$$P(r = k) = \{1 + \exp(-a(\theta - b_{k-1}))\}^{-1} - \{1 + \exp(-a(\theta - b_k))\}^{-1}$$
$$= p^*(k-1) - p^*(k)$$

where a is the slope or discrimination parameter,

b_k is the threshold or difficulty level for response category k

and $p^*(0) = 1$

$p^*(m) = 0.$

This model reduces to the two-parameter logistic model for a dichotomously scored items. Note that the 1.7 scaling factor is not present in the model.

Each MC item was dichotomously scored (0 incorrect, 1 correct) and each essay rating was fit with four graded categories. Nonresponses were coded as missing data. In the calibration runs the first GP and LU ratings only were used for both Essay 1 and Essay 2. Both ratings were not used together because it was thought that two purpose- or two language-usage ratings of the same essay would be highly dependent.

Three graded response calibrations were performed using MULTILOG (Thissen, 1985). One calibration run was done with the 72 MC items and the four essay ratings combined. A second run was done using only the 72 MC items, and the third run was done only using the four essay ratings. The purpose of the combined run was to provide a common scale (reference composite) for comparing MC and essay item parameters and item information. By comparing the item parameter estimates from the calibration of the 72 items with the item parameter estimates from the calibration of the combined sample, the approximate angle between the MC and the MC/Essay reference composite

could be determined. Likewise by examining the shift in item parameter estimates when the essay items were calibrated by themselves with the estimates of the combined sample, the approximate angle between the essay test and the MC/Essay reference composite also could be estimated.

Results

The first four moments of the MC content-score distribution, proportion-correct and biserial correlations averaged over number of items are shown in Table 1. Moments were also computed for the four reported essay score ratings. The most difficult MC content items were those in the OR category, and the most discriminating items, on average, belonged to the SL and G categories.

The MC test appears to be highly speeded. Only 90% of the examinees responded to item 58. This percentage continued to decrease with 78% of the examinees responding to the last four items.

Insert Table 1 about here

The correlation matrix of the six MC content scores with the four reported scores is shown in Table 2. The lower half of the matrix contains the uncorrected Pearson product-moment coefficients, reliabiility coefficients are located along the diagonal, and above the diagonal are the correlations corrected for unreliability. The reliability coefficients for the essays were computed by correlating the reported ratings for each essay.

Insert Table 2 about here

The corrected correlation matrix was factor analyzed by principal axes using squared multiple correlation as commonality estimates. A clear, two-factor solution was obtained with MC content areas loading highly on one factor and the essay ratings loading highly on the second. The obtained solution was rotated obliquely. The rotated factor loadings and eigenvalues are reported in Table 3.

Insert Table 3 about here

All three of the MULTILOG calibrations converged in less than 25 cycles. To examine the fit of the model, the inter-item correlations predicted by the graded response model (expected inter-item correlations) were computed and compared with the observed values. These correlations were averaged for three distinct groups: those involving pairs of MC items only, those involving pairs of essay ratings only, and those involving both an MC item and an essay rating. The observed and expected correlations are reported in Table 4. Both the within MC inter-item and within essay inter-rating correlations were underfit by the model. The MC/essay crossed correlations were reproduced more accurately. This pattern of mean residuals is consistent with the hypothesis that the combined MC and essay data are multidimensional.

Insert Table 4 about here

The means and standard deviation of the discrimination and difficulty parameters for each content area and essay rating type are displayed in Table 5 for the combined sample calibration. These results correlated quite highly with the classical item statistics, $r = -.93$ between item p-values and

IRT difficulty parameter estimates, and r = .85 between the biserial correlations and the IRT discrimination parameter estimates. The most discriminating content items were the SL items ($\hat{a}$ = .87), the most difficult items were the OR items, $\hat{b}$ = .14. Both the PS and LU essay ratings had much higher $\hat{a}$ values than any MC content item average, .90 and 1.45, respectively.

Insert Table 5 about here

One interesting analysis was to determine the angular difference between the MC trait, the essay trait, and the MC/Essay reference composite. The purpose of this is to get a rough idea of how much decrease in discrimination can be expected when MC items and essay items are mapped onto the same reference composite. The arc cosine of the ratio of the mean discrimination of the MC items calibrated with and without the essay ratings provides an approximation of the angle between the traits measured by the MC items and the MC/Essay reference composite. This was approximately 12°. A similar computation for the essay ratings revealed the angle between the traits measured by the combined essay ratings and the MC/Essay reference composite to be about 50° (48° for LU ratings and 53° for GP ratings.) The angle between the traits measured by the MC items and the traits measured by the combined essay ratings, 62°, can be compared to the arc cosine of the correlation between the total MC score and the average essay rating. This correlation was .53, suggesting that the angle between the essay and MC composites was approximately 58°.

IRT information functions for the individual MC content areas, and the essay rating type were computed and averaged over the number of items or number of ratings. These values are displayed in Table 6. These results,

which parallel the reliability and calibration results, reveal the amount information the essay ratings are providing in relation to their MC counterparts on the MC/Essay reference composite. The LV essay rating was the most informative of any MC content area or essay rating across the entire ability range. The OR content area provided the least amount of information of any MC content area or essay rating. The G content category was the most informative of the MC content categories and peaked at a theta of -1.0.

Insert Table 6 about here

The relative information was further analyzed by constructing two information plots. The first plot, shown in Figure 2, displays the information provided by each MC content area and the essay rating type averaged over the number of items in each content area or ratings for each essay type. Compared to the average MC content area information, the LU essay ratings provided more information over the entire ability scale.

Insert Figure 2 about here

The second plot shows the ratio of the total test information of the MC items plus the essay rating to the total test information of the MC items alone. This relative information plot shows the added information in terms of MC items that can be gained by using the essay rating to supplement the MC test results. Clearly the essay prompts and scoring protocols provide more information at the upper end of the ability scale, approximately the equivalent of about 40 MC items.

Insert Figure 3 about here

Discussion and Conclusion

The purpose of this study was to determine if the CAAP essay and MC tests were measuring the same abilities, and secondly to compare the two tests on the amount of common information they provided. Because CAAP is in the development phase, both of these questions are important.

Pertaining to the first issue, factor analytic results strongly suggest that the essay and MC tests are measuring different abilities. The rotated factor loadings indicated that when response data from the two measures of writing were combined, one factor was clearly marked by the MC content areas, while the second was dominated by the essay GP and LU ratings. These findings coincide with those of Ward et al. (1980), and Ackerman and Smith (1988), who also determined that essay ratings can provide unique information.

Results also indicate that the two types of essay ratings may be measuring different writing skills. This is suggested by the fact that the LU ratings correlated higher with each MC content area than did the GP ratings. Further evidence is that the angular difference between the MC/Essay reference composite and the LU reference composite is less than the angular difference between the MC/Essay reference composite and the composite for GP. That is, the LU ratings appear to be measuring traits more similar to the MC test than are those being measured by GP ratings.

Regarding the second issue, the amount of common information provided by both measures, the results are somewhat surprising. Plots of the added information gained by combining the essay ratings to the MC test results (at the upper end of the ability scale) show the increases in information would be

as high as 40 MC items. The essay ratings are highly reliable and therefore highly discriminating. This is in spite of the fact that the $\hat{a}$ values are underestimates due to the mapping onto the MC/Essay reference composite.

One point of concern is that typically MC items tend to have guessing and this cannot be handled in the graded response model calibrated with MULTILOG. An approach that could be taken would be to estimate the MC items with a three parameter IRT model, fix the $\hat{a}$ and $\hat{b}$ parameters for these items, and then estimate the essay ratings. In a sense, this would be directly placing the essay items on the MC scale.

It should also be noted that the IRT calibration model did not contain the 1.7 scaling factor, thus the discrimination parameter estimates and information values are probably overestimates of their true values.

The next step in evaluating the relative worth of the two CAAP writing measures would be to obtain a criterion score, (i.e., college English GPA) and see how much more predictive power one instrument adds beyond the predictive power of the other instrument. This type of analysis would answer the question about the relative worth of the unique information each test provides.

References

Ackerman, T.A. and Smith, P.L. (1988). A comparison of the information provided by essay, multiple-choice and free-response writing tests. Applied Psychological Measurement, 12, 117-128.

Braddock, R., Lloyd-Jones, R., and Schoer, L. (1963). Research in written composition. Champaign, IL: National Council of Teachers of English.

Breland, H.M. (1977). A study of college English placement and the test of standard written English. Princeton, NJ: Educational Testing Service.

Cooper, C.R. and Odell, L. (1977). Consideration of sound in composing process of published writers. Research in Teaching English, 10, 103-115.

Dorans, N.J. and Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22, 249-262.

Samejima, F. (1969). Estimation of latent ability using a response patter of graded scores. Psychometrika Monograph Supplement, 4, 2, 17.

Thissen, D. (1985). MULTILOG, Version 4.0 User's Guide. Lawrence, KS: University of Kansas.

Tucker, L.R. and Finkbeiner, C.T. (1981). Transformation of factors by artificial personal probability functions. Educational Testing Service Research Report.

Wang, M.M. (April, 1986). Fitting a unidimensional model to multidimensional item response data. Paper presented at the ONR Contractors Conference, Catlinburg, TN.

Ward, W. C., Frederiksen, N. and Carlson, S.B. (1980). Construct validity of free-response and machine-scorable forms of a test. Journal of Educational Measurement, 17, 11-29.

Yen, W. M. (1985). Increasing item complexity: a possible cause of scale shrinkage for unidimensional item response theory. Psychometrika, 50, 399-410.

Table 1

The mean, standard deviation skewness, kurtosis, average p-value and average biserial correlation for each MC content area and essay rating

Content Area	n	Mean	STD	Skewness	Kurtosis	$\bar{p}/\sigma_p$	$\bar{r}_{bis}/\sigma_{r_{bis}}$
G	8	5.07	1.89	-.40	-.78	.64/.14	.50/.16
OR	10	4.93	2.16	-.04	-.53	.49/.15	.47/.12
SS	18	10.00	3.59	-.07	-.95	.56/.15	.48/.11
P	6	3.19	1.53	.01	-.76	.53/.20	.45/.09
SL	14	8.79	3.08	-.27	-.75	.63/.12	.53/.08
SA	16	9.05	3.19	-.30	-.80	.57/.17	.48/.15
TOTAL	72	41.04	12.76	-.14	-.69	.57/.15	.49/.12
ESSAY 1							
GP	2	2.60	.75	-.22	-.34		
LU	2	2.69	.57	-.12	.31		
ESSAY 2							
GP	2	2.14	.82	.16	-.78		
LU	2	2.65	.60	-.34	.18		

Table 2

Uncorrected correlations (below the diagonal) and corrected correlations (above the diagonal) between MC content categories and essay rating scores

	MC Test						Essay			
	G	OR	SS	P	SA	SL	GP1	LU1	GP2	LU2
G	**.60**	.96	.86	.68	.94	.91	.50	.58	.44	.53
OR	.56	**.57**	.96	.86	1.00	1.00	.55	.66	.57	.66
SS	.57	.62	**.73**	1.00	1.00	1.00	.72	.80	.70	.73
P	.37	.46	.62	**.50**	.85	.89	.61	.75	.56	.68
SA	.61	.65	.72	.50	**.70**	1.00	.67	.74	.65	.73
SL	.60	.70	.73	.54	.72	**.73**	.59	.70	.62	.68
GP1	.27	.29	.43	.30	.39	.35	**.49**	.89	1.00	.78
LU1	.36	.40	.55	.43	.50	.48	.50	**.65**	.69	1.00
GP2	.24	.30	.42	.28	.38	.37	.49	.39	**.49**	.80
LU2	.33	.40	.50	.39	.49	.47	.44	.65	.45	**.65**

Note: Reliability coefficients are located along the diagonal.

Table 3

Rotated factor loadings for the six MC content areas and the four essay rating scores

Content Area	1	2	3	4
G	1.06	-.19		
OR	1.07	-.10		
SS	.85	.21		
P	.78	.17		
SA	.94	.09		
SL	1.00	-.01		
GP1	-.09	1.01		
LU1	.22	.74		
GP2	-.08	.96		
LU2	.12	.81		
Eigenvalues	7.88	1.20	.33	.24

Note: Factors 1 and 2 were correlated .69.

Table 4

Observed and expected mean inter-item correlations for MC items, MC items/essay ratings, and essay ratings

	Observed			Expected	
	MC	Essay		MC	Essay
MC	.13	.15	MC	.10	.15
Essay		.42	Essay		.39

Table 5

The mean and standard deviation of estimated discrimination and difficulty parameters for the six MC categories and the GP and LU essay ratings

Content Area	n	$\hat{a}$		$\hat{b}$		$\hat{b}_1$		$\hat{b}_2$		$\hat{b}_3$	
		$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ	$\bar{X}$	σ
G	8	.72	.39	-1.25	.87						
SS	18	.78	.24	-.42	1.05						
OR	10	.68	.22	-.14	.93						
P	6	.78	.22	-.36	1.28						
SL	14	.87	.19	-.95	.61						
SA	16	.76	.34	-.45	.95						
GP	2	.90	.01			-2.14	1.09	.19	.84	3.19	.71
LU	2	1.45	.01			-3.37	.48	-.49	0.06	2.55	.04

Table 6

Average IRT information values for the six MC content areas and the GP and LU essay ratings for selected theta values

Content Area	n	Theta						
		-3.0	-2.0	-1.0	0.0	1.0	2.0	3.0
C	(8)	.23	.25	.26	.22	.14	.14	.14
SS	(18)	.07	.12	.15	.14	.10	.06	.03
OR	(10)	.05	.08	.11	.12	.10	.06	.04
P	(6)	.05	.09	.14	.15	.12	.08	.04
SL	(14)	.09	.15	.18	.16	.10	.05	.02
SA	(16)	.08	.13	.16	.13	.08	.05	.03
GP	(2)	.17	.21	.23	.23	.22	.21	.21
LU	(2)	.91	1.00	1.05	.84	.52	.57	.48

Note: These information values are based upon item parameter estimates obtained from the "combined" calibration run.

Figure Captions

Figure 1. The mapping of an item onto the test reference composite.

Figure 2. The average item information of six MC content areas and GP and LU essay ratings.

Figure 3. Effective test length resulting from adding Essays to MC items.

Figure 1

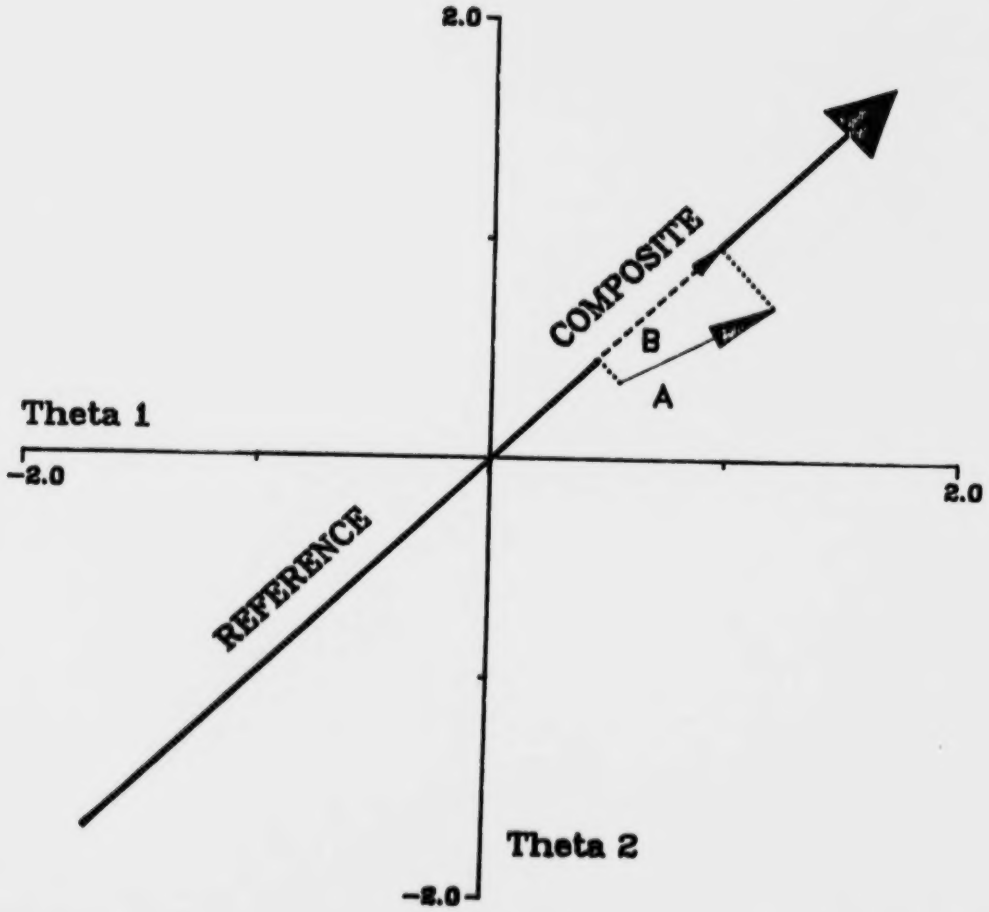


Figure 2

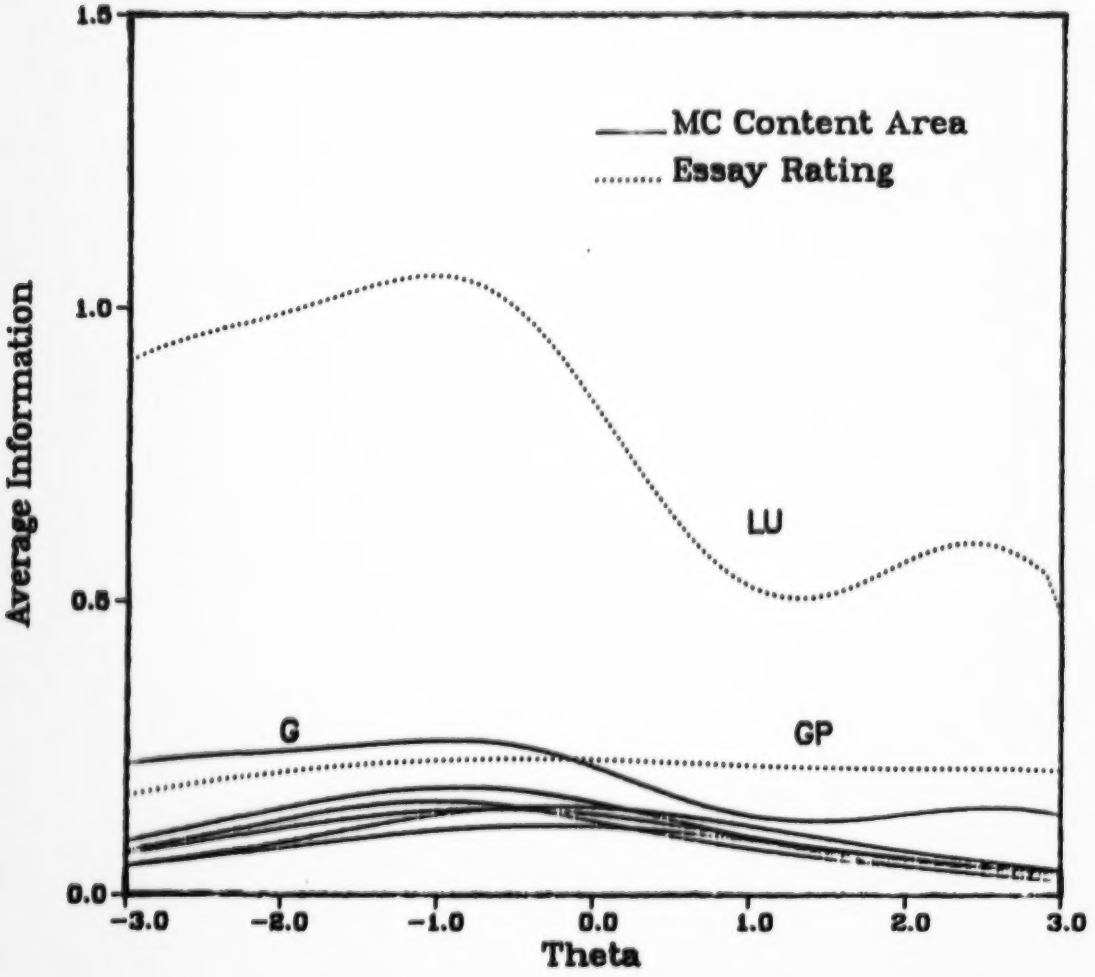
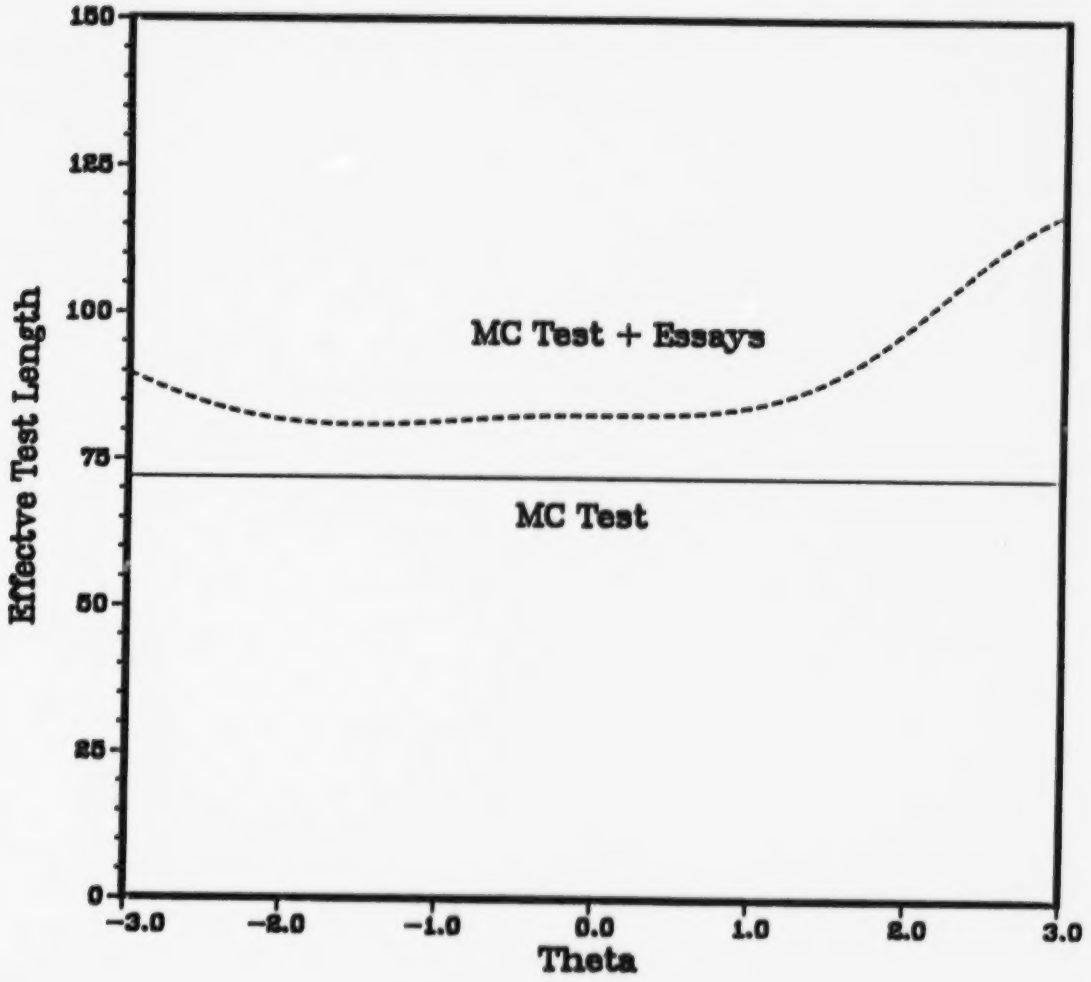

1.5
1.0
0.5
0.0
Average Information
MC Content Area
Essay Rating
LU
G
GP
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

Figure 3


150
125
100
75
50
25
0
Effectve Test Length
MC Test + Essays
MC Test
-3.0
-2.0
-1.0
0.0
1.0
2.0
3.0
Theta

Appendix

Example Essay Prompt

Your college administration is considering whether or not there should be a physical education requirement for undergraduates. The administration has asked students for their views on the issue and has announced that its final decision will be based on how such a requirement would affect the overall educational mission of the college. Write a letter to the administration arguing whether or not there should be a physical education requirement for undergraduates at your college.

(Do not concern yourself with letter formatting; simply begin your letter, "Dear Administration.")

GP Score Point Descriptions

4 Elaborated appropriate argument. These papers take a position on the issue defined in the prompt and support that position with an elaborated argument of appropriate reasons. The argument's main ideas are logically connected and thoroughly developed. These papers clearly recognize the grounds upon which the issue will be resolved and the argument clearly focuses on those grounds.

3 Appropriate argument. These papers take a position on the issue defined in the prompt and support that position with an argument consisting of several appropriate reasons. The argument's main ideas are logically connected and one or two may be somewhat developed, but the argument as a whole does not constitute an elaborated argument. These papers clearly recognize the grounds upon which the issue will be resolved and the argument generally focuses on those grounds.

2 Brief but appropriate argument. These papers take a position on the issue defined in the prompt and support that position with a brief argument of appropriate but undeveloped reasons. These papers clearly recognize the grounds uponw hich the issue will be resolved, but the argument either does not focus on those grounds (a number of reasons, two or more appropriate, but most inappropriate) or is so brief as to offer only a position and a couple of undeveloped appropriate reasons.

1 No appropriate argument. These papers take a position on the issue defined in the prompt but offer only one undeveloped appropriate reason in support of that position. Of these papers take a position but do not support that position with any appropriate reasons. Or these papers do not take a clear position on the issue. These papers may not recognize the grounds upon which the issue will be resolved, or they recognize the grounds but simply dismiss them.

OT Off task. These papers are not ratable because they are totally irrelevant to the prompt or refuse to engage the task.

I Illegible. These papers are not ratable because the writing is illegible.

NE Not English. These papers are not ratable because they are written in a language other than English.

NR No response. These papers are not ratable because they do not respond to the prompt at all.